Learning Environ Res (2016) 19:335–357
DOI 10.1007/s10984-016-9215-8

ORIGINAL PAPER

Observations and student perceptions of the quality of preservice teachers' teaching behaviour: construct representation and predictive quality

Ridwan Maulana[1] · **Michelle Helms-Lorenz**[1]

Received: 14 July 2014 / Accepted: 30 May 2015 / Published online: 21 July 2016

Abstract Observations and student perceptions are recognised as important tools for examining teaching behaviour, but little is known about whether both perspectives share similar construct representations and how both perspectives link with student academic outcomes. The present study compared the construct representation of preservice teachers' teaching behaviour as perceived by trained teacher observers and students. It also examined the predictive power of both measures of teaching behaviour for student academic engagement. The theoretical framework of teaching behaviour used in this study is based on evidence-based research derived from empirical teacher effectiveness research. The study was part of a national project that included 2164 students and 108 teachers in The Netherlands. Results suggest that, although observations and student perceptions of teaching behaviour shared similar theoretical considerations, the construct representations seemed to differ to some extent. Furthermore, although both perspectives are significant predictors of student academic engagement, student perceptions appeared to be more predictive of their perceived academic engagement than observations. Implications for research on learning environments and teacher education are discussed.

Keywords Construct representation · Peer teacher ratings · Predictive quality · Student perceptions · Teaching behaviour

Introduction

Research confirms that classroom social climates, in terms of teaching behaviours, are important predictors of student success at schools (Maulana 2012; Van de Grift 2007). However, how best to examine teaching behaviour remains inconclusive. Learning

✉ Ridwan Maulana
r.maulana@rug.nl

[1] Department of Teacher Education, University of Groningen, Grote Kruisstraat 2/1, 9712, TS, Groningen, The Netherlands

environments research recognises that teachers, students and observers have access to different features of learning environments (Seidel and Shavelson 2007). This suggests that the three sources of measurement are important sources for studying learning environments. Nevertheless, the history of learning environments research is strongly characterised by the construction and application of instruments from student and teacher views (Cavanagh and Romanoski 2006). Comparing teaching behaviour using observations and student or teacher perspectives in learning environments research is rather limited. Hence, little is known about the degree to which the construct representation measured by observations and student surveys is comparable and how best to interpret the potential differences in perceptions.

As in many countries, there is a national initiative in The Netherlands to improve education by improving the quality of many educational levels (i.e. teacher education institutions, schools, teachers, contexts). Specifically, the improvement of the quality of teachers' classroom performance has been emphasised to be one of the major concerns (Ministerie van Onderwijs, Cultuur en Wetenschap 2012). In general, Dutch students' performance in international comparison studies (e.g. PISA, TIMSS) has been good, with a slight drop in (mathematics and English) performance observed in the last investigation (OECD 2010). The next step that the country would like to achieve is to upgrade the good performance to the excellent level. One important approach for achieving this is by developing methods for assessing teaching behaviour that can help teachers to develop their teaching behaviour in their zone of proximal development. The application of item response theory is promising for this purpose (Cavanagh and Waugh 2011; Maulana et al. 2015).

Comparisons between teacher and student surveys measuring teaching behaviour reveal rather low convergent patterns, as well as low predictive power for student outcomes (Kunter and Baumert 2006; Maulana et al. 2011). Kunter and Baumert (2006) argue that teacher and student perceptions of instructional practices are not merely different methodological approaches to examining the same components of teaching practices. Likewise, we anticipate that classroom observations by external observers and student surveys of teaching behaviour are not simply different methodological approaches as well. The methods might be tapping different representations of the meaning of 'good' teacher behaviour. Both parties could attribute similar or different meanings to certain teacher behaviour.

In this study, we aimed to compare the construct representation of the teaching behaviour of preservice teachers as perceived by observers and students by means of classical test theory and item response theory (IRT). Specifically, Rasch modelling was applied. Next, we examined the predictive power of classroom observations and student surveys of teaching behaviour on student academic engagement. Combining classical test theory and Rasch modelling together offers potential insight into the current knowledge base about the reliability, validity and usefulness of different methods for measuring classroom practices.

Theoretical framework

Observation and survey as indicators of teaching behaviour

One dilemma with regard to evaluating (the effectiveness of) teaching behaviour deals with selecting techniques that are powerful (enough) to reflect 'real' classroom practice and, at

the same time, are predictive of various student outcomes. To date, there are (at least) three common perspectives for measuring teaching behaviour: student self-report surveys, teacher self-report surveys, and external classroom observations (Lawrenz et al. 2003). Compared with observational measures, student and teacher surveys are mostly preferred because these methods are considered more economical (cost-effective) than observations (Fraser 1991, 2012). Besides, both student and teacher perceptions are based on their (classroom) experiences over time, not merely from a single or limited number of observations. Therefore, student evaluation of teaching has been one of the most widely-used indicators of teacher effectiveness and educational quality (Spooren et al. 2012).

Furthermore, results of student surveys (but not teacher surveys) in the class can be aggregated to the class level. De Jong and Westerhof (2001) showed that the quality of aggregated class-level student perceptions is comparable to that of observation measures. In fact, what students actually perceive could be more important than what outsiders observe because student perceptions steer their own learning behaviour, based on their own insights. Indeed, studies indicate that student perceptions are mostly more predictive of student outcomes compared to observations (De Jong and Westerhof 2001; Seidel and Shavelson 2007) and teacher perceptions (Scantlebury et al. 2001). Additionally, teachers tend to overrate their own teaching behaviour and underestimate their judgements regarding negative teaching behaviour (Maulana et al. 2012a, b).

However, student and teacher perceptions can be affected by how they perceive that the data will be used—the so-called social desirability effect (Van de Grift et al. 1997). Therefore, teachers might respond to questionnaires in a more socially-accepted way than the real practice reveals. When possible, social desirability effects should be controlled for in survey data. In addition, many external factors (i.e. teaching subject, teacher gender, student gender) can influence the way in which respondents perceive teaching behaviour that could reduce the objectivity of this technique (Aleamoni 1981; Maulana 2012a, b). Therefore, results from self-report survey data should be interpreted with care and should not be over-extrapolated (Saljo 1997).

Classroom observations, compared with student and teacher surveys, are considered as the most objective method of measuring teaching practices (Worthen et al. 1997). This method is recognised as an important procedure in the teacher training process (Lasagabaster and Sierrra 2011). Classroom observations allow judgements about what is happening in the classroom, and these judgments are assumed to be 'free' from the influence of the first (i.e. students) and the second (i.e. teachers) parties (Lawrenz et al. 2003). Nevertheless, the presence of observers can influence teachers' behaviour as well (De Jong and Westerhof 2001). Therefore, classrooms need to adapt to the presence of the observers, before the real observation is conducted, in order to minimise the so-called observer effects.

Given that both observations and student surveys have strengths and weaknesses, both methods should be seen as complementary ways to measure teaching behaviour (triangulation). Triangulation has also been viewed as a technique to ensure the validity and reliability of instruments measuring complex classroom practices (Denzin 1997). However, Riggin (1997) argued that triangulation can result in either complementary or conflicting findings. Besides, using several methods to collect data for measuring teaching behaviour is not efficient in terms of time and practical considerations. Consistent with Lawrenz et al. (2003), we believe that it would be beneficial for classroom researchers to determine if a particular method of data gathering would generally be useful for measuring teaching behaviour because the literature reviewed earlier indicates the promising quality of student surveys relative to teacher surveys for measuring teaching behaviour, it is logical that, after

a valid and reliable observation instrument is available, constructing student surveys would be the next logical step.

Construct representation of teaching behaviour

A major challenge in educational research is that there is insufficient evidence regarding convergence between observations and student perceptions of teaching behaviour. The concept of convergence is related to a construct's representation and significance. As an aspect of the internal validity, construct representation involves "the meaning of the construct, in terms of the processes, strategies, and knowledge that are directly involved in performance (Embretson 2007, p. 15). Low correlations between observations and student perceptions of teaching behaviour are indicative of differences in the construct representation by the parties involved. Additionally, differing patterns of item difficulties within an instrument measuring teaching behaviour can be a reflection of differences in the underlying mechanisms of teaching performance (Embretson 2007).

Construct significance deals with the predictive power of a construct to another measure or other external criteria (external validity). A construct is considered significant when it shows a significant relation to other (outcome) variables. For example, teaching behaviour can be conceived of as a significant construct if it relates to academic engagement of students: the better the quality of teaching behaviour, the higher the level of academic engagement of students, and vice versa.

Teaching expertise and teaching behaviour

Teaching behaviour is a complex concept and multidimensional in nature (Shuell 1996). We acknowledge different definitions of effective teacher behaviour as summarised by Ko and Sammons (2013) and we endorse the definition focusing the effectiveness of observable behaviours as seen in the classroom during a regular lesson. In the present study, we focused on observable classroom behaviour that has been shown to have an impact on student achievement. Choosing a narrow definition of teacher effectiveness can be cost-efficient when the aim is to measure longitudinal improvement in effective classroom behaviour. The instruments compared in this study are designed to measure the impact of interventions aimed at improving (novice) teacher behaviour development.

According to reviews of research on the relationships between the basic characteristics of teaching and academic outcomes, there are several observable teaching behaviour components that are closely connected to the effectiveness of teaching: *creating a safe and stimulating learning climate, exhibiting efficient classroom management, displaying clear instruction, activating learning, employing adaptive teaching* and *implementing teaching and learning strategies* (see Creemers 1994; Fraser 2007, 2012; Sammons et al. 1995; Seidel and Shavelson 2007; Van de Grift 2007; Wubbels and Brekelmans 2012 for reviews).

Early studies on teaching expertise revealed qualitative differences between novice and expert teachers: novice teachers have more difficulty when asked to interpret classroom phenomena (Sabers et al. 1991); experts exhibit smoother routines (Berliner 1988), are more accurate in their assumptions and hypotheses, and are more emotionally involved in their work than novices (Berliner 1988). The work of Day (2007) revealed valuable insights into factors influencing the development of teachers' professional identity and other psychological variables. However, these descriptions and insights are not easy to operationalise when the aim is to reveal (small) developmental steps in (short-span) longitudinal studies.

Valuable insights have been gathered with respect to subsets of teacher behaviour, such as interpersonal relationships by Wubbels and Brekelmans (2005). A small number of studies incorporate multi-behavioural facets in concert (e.g. Kyriakides et al. 2009; Maulana et al. 2015; Sammons and Ko 2008; Van de Grift et al. 2011, 2014), thus allowing testing for the unidimensionality of the construct 'teaching behaviour' and the possibility of ranking items by item difficulty. These behavioural instruments with a broader spectrum of observable behaviour are more suitable for measuring the impact of interventions aimed at accelerating teaching behaviour when teachers of various proficiency levels are involved.

Recent studies in primary and secondary schools revealed that teachers with 15–20 years of experience show the highest average performance levels of effective teaching behaviour compared with less- and more-experienced teachers (Van de Grift and Helms-Lorenz 2012). Among the teachers, preservice teachers show the lowest performance levels, followed by beginning teachers. The ascending trend of teaching behaviour performance reaches a peak after 15–20 years of experience, after which it starts to descend fairly during in the period of 20–30 years of experience. After this period, teaching performance begins to rise slightly. The pattern of this teaching behaviour performance has been replicated in (cross-sectional) studies among various European countries (Van de Grift 2007). From this study, we conclude that many years of experience seem to be necessary to reveal highly-effective teaching behaviour. It would be beneficial for educational outcomes to have teachers with less experience displaying effective teaching behaviour. We argue that this objective can be accomplished by increasing the level of performance of preservice and beginning teachers. These group of teachers have great impact on education because they stay in the teaching profession for the longest period of time. To achieve this goal, the use of valid and reliable instruments for evaluating teaching behaviour becomes essential.

Teaching behaviour and student engagement

Past research using teacher and student surveys shows that the quality of teaching behaviour is associated with students' academic engagement (Anderson et al. 2004; Furrer and Skinner 2003; Klem and Connell 2004; Patrick and Ryan 2007; Roorda et al. 2011). Academic engagement, as in the present study, is conceptualised as the extent to which students are psychologically and behaviourally engaged in academic tasks (Appleton et al. 2006; Van de Grift 2007). Students who have positive perceptions about teacher support show good behavioural engagement with schooling and obtain high learning achievements (Woolley and Bowen 2007). Additionally, student engagement in classroom activities serves as a starting point for good academic grades (Finn 1989). Hence, student engagement can be seen as an indicator of classroom effectiveness and a mediator between classroom processes and student achievements (Virtanen et al. 2013).

Studies of the relationship between student engagement and classroom quality have been typically undertaken using student and teacher surveys. In secondary education particularly, the use of classroom observation to examine the relationship between classroom quality and student engagement remains relatively scarce (Virtanen et al. 2013). Research shows that observed teacher support in the classroom is a strong predictor of student engagement and achievement as well (Allen et al. 2013; Cornelius-White 2007; Roorda et al. 2011).

Based on the literature reviewed earlier, our study was designed to answer the following research questions:

1. How does the construct representation of observations compare with student perceptions of teaching behaviour?
2. How does the predictive quality of observations compare with student perceptions of teaching behaviour? More particularly, are observations and student perceptions significant predictors of academic engagement?

We hypothesise that the construct representation of teaching behaviour is different between observations and student perceptions. We also hypothesise that both observations and student perceptions of teaching behaviour are significant predictors of student engagement. Students' perception of teaching behaviour is a stronger predictor of their academic engagement compared to observations.

An accurate, reliable and valid measure of teaching behaviour would be a useful tool in teacher assessment for measuring individual progression (micro level) but also to measure the progression of the nationwide (preservice) teacher taskforce (macro level). Additionally, the instrument can be useful in the scientific community to measure the effects of interventions designed to accelerate teacher development, thus promoting the improvement of educational outcomes.

Methods

The data reported are part of a national, longitudinal study of the relationship between the development of the quality of teaching behaviour of preservice teachers and the preparation route of teacher education for Dutch secondary education. For each class participating in the study, students completed the questionnaire and the majority of their teachers were observed by peer teachers who focused on various aspects of teaching behaviour displayed by preservice teachers.

Sample

A total of 108 preservice teachers participated in this study. Of the participating teachers, 69 were female and 39 were male. The average class size was 22.8 (SD = 5.4). Teaching experience ranged between 0 and 4 years (mean = 2.1, SD = 0.5). A small proportion of classes/teachers of the participating students were not observed. Hence, the sample for the student survey data was relatively larger than for the observation data. The study included a nationally-representative sample of 2164 students of grade 9–13 in 98 classes from 44 secondary schools throughout the Netherlands. Students were distributed as follows across grades: 30 % of the students were in grade 9, 32 % in grade 10, 21 % in grade 11, 14 % in grade 12 and 3 % in grade 13. All schools were recruited based upon voluntary participation, and all the schools in the Netherlands were approached. The age of the students ranged from 11 to 18 years (mean 16 years) and 53 % of them were girls.

Measure

Observations

In order to identify the teaching behaviour of preservice teachers, the observation instrument originally developed for the International Comparative Analysis of Learning and Teaching (ICALT, Van de Grift 2007) was used. This observation instrument was

validated across several European countries including the Netherlands, Belgium, Germany and England. The observation instrument was found to be reliable and valid for measuring teaching behaviour (Van de Grift 2007; Van der Grift and Van der Wal 2010). Over the last couple years, this observation instrument has been used in the Netherlands for teacher training and research purposes. This instrument shows construct overlap with observation instruments developed by Danielson (2013) and Pianta and Hamre (2009).[1]

The instrument consists of 32 items measuring six components of teaching behaviour, namely (1) *Safe and stimulating learning climate* (four items), (2) *efficient classroom management* (four items), (3) *Clear instruction* (seven items), (4) *Activating learning* (seven items), (5) *Adaptation of teaching* (four items) and (6) *Teaching–learning strategies* (six items, see Appendix 1 for an example of items). The behaviours were coded on a four-point Likert scale representing the presence of the observed teaching behaviour, ranging from 1 (predominantly weak) to 4 (predominantly strong). Higher scores within each scale and across the dimension of teaching behaviour represent more effective teaching behaviour, and vice versa.

Every observer observed preservice teachers' classroom practices (natural classroom observation) using the observation instrument independently. Real-time coding (in lessons) was used. Observers sat at the back of the classroom during the entire lesson (whole-lesson coding). They focused first on examples of good practices (low-inference indicators). These low-inference indicators help observers to rate the high-inference indicators.

Training of observers

All trained observers in this study were experienced teachers (>5 years of experience). Not only were observers experienced in teaching, they also were experienced with coaching preservice teachers. The majority of the observers were school coaches and school mentors. Some observers observed multiple preservice teachers. They observed preservice teachers teaching in their schools. All observers were trained prior to performing classroom observations, which were organised in sessions of 5–12 participants. Four trainers involved were lecturers of the department of teacher education. Background information about the items was presented, based upon studies showing evidence of effectiveness of teacher behaviour. The scoring procedure for behaviour was explained. The observation instrument was studied and questions concerning comprehension and formulation were answered. A video fragment of a lesson of 15 min was shown. Trainees judged the behaviour using the observation instrument. The trainers requested all the participants to reveal their judgements. Differences and similarities were discussed and defended with the aim of reaching consensus. A second video fragment was presented following the same procedure. The forms were collected to determine the inter-rater reliability. Observer agreement for all teaching behaviour scales was reached at an above-satisfactory level (inter-rater reliability >0.70).

Student questionnaire

As part of the national project, a student questionnaire, which is theoretically consistent with the observation instrument, was developed to capture student perceptions of their teachers' teaching behaviour. The items of the student questionnaire were developed through the following steps: (1) adjusting items of the ICALT observation instrument in terms of language use and formulation in order to meet the comprehension level of

[1] For a comparison, see Maulana et al. (2014).

secondary-school students about the measured constructs; (2) studying the literature about effective teaching behaviour in order to generate ideas about additional items; and (3) generating additional items based upon the reviewed literature on effective teaching behaviours and the existing scales of the ICALT observation instrument. To ensure that all items in the student questionnaire were qualitatively comparable to the items and scales of the observation instrument, face validity was checked by the authors together with the original developer of the ICALT instrument.

Compared with the observation instrument, which is more holistic regarding its wording and formulation because it is targeted at trained observers who are knowledgeable about effective teaching constructs and behaviour, the student questionnaire is more concrete and specific and represents items that are reflective of the same aspects of effective teaching behaviour that are suitable for tapping student perceptions in secondary education. The first draft of the questionnaire consisted of 71 items. After performing initial reliability analyses, it was found that eight items contributed negatively to the scale reliability (two were reverse-scored items). After conducting face validity for these eight items, we decided to exclude the items from the questionnaire and used the 64-item version for further analyses. The number of items within each scale in the 64-items version is: (1) Safe and stimulating learning climate (12 items), (2) Efficient classroom management (eight items), (3) Clear instruction (11 items), (4) Activating learning (17 items), (5) Adaptation of teaching (six items) and (6) Teaching–learning strategies (ten items, see Appendix 2 for an example). All items were answered on a four-point Likert scale ranging from 1 (Completely disagree) to 4 (Completely agree).

Academic engagement

Student self-report of academic engagement measure was used as an external criterion to examine the predictive quality of teaching behaviour. The measure was based on a scale developed by Van de Grift (2007), which is conceptually consistent with that of Maulana (2012a, b), which emphasises psychological and behavioural engagement. The scale consists of six items provided on a four-point response scale ranging from 1 (completely not true) to 4 (completely true). Examples of items are "I participate well during the lesson", "I do my best during the lesson" and "I pay attention during the lesson". The internal consistency of the scale is very good (Cronbachs' $\alpha = 0.88$).

Data-analytic approach

Prior to answering the research questions, we conducted preliminary psychometric analyses including exploratory factor analyses (EFA), reliability analyses, correlational analyses and intra-class correlations on peer teacher and student ratings. The results of the preliminary analyses were important for determining subsequent analyses. Smith (1996) showed that, when the data are dominated equally by uncorrelated factors, factor analysis is an appropriate method for examining the unidimensionality of a latent construct but, when the data are dominated by highly correlated factors, Rasch modelling is more suitable. Therefore, we conducted factor analysis as a preliminary examination of the data.

Furthermore, we performed Rasch modelling (with dichotomous responses)[2] in order to examine the construct representation of the observation instrument and the student

[2] Following Maulana et al. (2015), the original four-response category scores were transformed into dichotomous scores using the following criteria: 1 and 2 = 0, 3 and 4 = 1.

questionnaire on teaching behaviour (Research question 1). Rasch modelling offers advantages over classical test theory for studying teaching behaviour: "(1) producing linear, unidimensional scales; (2) requiring that data must fit the measurement model; (3) producing scale-free person measures; (4) producing sample-free item difficulties; (5) calculating standard errors; (6) estimating person measures and item difficulties on the same linear scale in standard units (logits); and (7) checking that the scoring system is being used logically and consistently. An additional reason for the increased use of Rasch measures is that they have been shown to be better than 'measures' based on classical test theory (just using the total score as the 'measure' on a set of items after and, most often, not even with Confirmatory Factor Analysis). Rasch measurement requires the researcher to design the items in a scale from easy to hard, but with certain conditions in mind. The conditions mean that the probability of answering positively must be related to the difference between the person measure (technically the person parameter) and the item difficulty (technically the item parameter)" (Cavanagh and Waugh 2011, p. xi).

The ordinary unidimensional Rasch model can be considered as a family of item response theory (IRT) which has more stringent assumptions compared to a classical test theory family. In order to satisfy the Rasch model, a set of items in the instrument should meet the following assumptions: (1) the items in a Rasch scale must have the same discriminatory power (assumption of parallelism of the item characteristic curves); (2) together, the items must measure one and the same latent skill in a homogeneous way (assumption of unidimensionality); and the response to one item may not influence the response to another, except for an influence that can be explained by the latent variable that is the measurement objective of the set of items (assumption of local stochastic independence) (Van de Grift and Van der Wal 2010).

Von Davier (1994) introduced an extension of the Rasch model, called the mixed Rasch model, which extends the ordinary unidimensional Rasch model by assuming that a non-homogeneous sample can be classified into Rasch-homogeneous subsamples. This means that the Rasch model might not hold for a given data set although it is expected to hold from theoretical points of view because the data consist of more than one homogeneous subgroup of subjects. The Rasch model holds within each of these subgroups, meaning that the items form a unidimensional construct within each subgroup, but that the subgroups reveal different rank orders of item difficulties. The model fit of the Rasch model was tested by comparing the one-subgroup (ordinary Rasch model) and the mixed Rasch model. Because we did not have a specific hypothesis about the number of the latent subgroups, we tested a number of latent subgroups until the fit of the model could not be improved. To select the best-fitting model, the fit statistic indicator called Consistent Akaike Information Criterion (CAIC; Bozdogan 1987)[3] was used. The lowest value on this criterion is indicative of the best model-data fit.

To evaluate the model-data fit within a model (i.e. in a Rasch model), the fit statistics of likelihood ratio (LR), Cressie-Read (CR), Pearson χ^2 and Freeman-Tukey (FT) were used. These fit statistics are derived from a parametric bootstrapping method which is a better way of evaluating models than information criteria. When the data are sparse (i.e. lots of possible response patterns, most of which are not observed), only C-R and Pearson χ^2 statistics are suitable to interpret (Von Davier and Rost 1994).

[3] Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC) and Consistent Akaike's Information Criterion (CAIC) can be used to compare models. However, AIC risks an overestimation when the sample is large ($N > 100$; Tu and Xu 2012).

Next, we performed multilevel analyses (with class as level 2 and student as level 1) to investigate the predictive quality of peer teacher and student ratings of teaching behaviour on students' academic engagement (Research question 2). Based on past research showing that teaching subject, teacher gender and student gender affect student engagement and teaching behaviour (Maulana 2012a, b; Opdenakker et al. 2012), these variables were entered as control variables. All significant results under the 95 % confidence interval were retained. We used SPSS version 20, WINMIRA (Von Davier 1994), Mplus (Muthen and Muthen 1999) and MLwiN (Rasbash et al. 2005) to analyse the data.

Results

Construct representation observations and student perceptions of teaching behaviour

Preliminary analysis

We performed factor analyses using principal component analysis on the observation data. Results indicated that a six-factor solution could be extracted. The six components accounted for 68 % of the variance. The first component accounted for 42 % of the variance, the second for 8 %, the third for 6 % and the remaining for about 4 % each. Reliability analyses showed that all scales had good internal consistencies (see Table 1), ranging from 0.72 (Adaptation of teaching) to 0.89 (Teaching learning strategy). Intra-class correlations between scales ranged between 0.37 (Activating learning) and 0.60 (Teaching–learning strategy), indicating that a significant amount of variance could be found at the teacher/class level. This means that the observation scales could distinguish between teacher/class differences in teaching behaviour. Moreover, mean inter-scale correlations ranged from 0.59 (Safe and stimulating learning climate) to 0.69 (Clear instruction and Activating learning). This indicates that, although there was an overlap among teaching behaviour components, the scales measured distinct aspects of teaching behaviour. A much larger variance explained by the first component than the remaining ones and high inter-scale correlations suggest preliminary support for the unidimensional construct of teaching behaviour as measured by the observation instrument.

Furthermore, consistent with the number of hypothesised scales as evident in the observation instrument, results of factor analyses of the student data showed that the items loaded onto six components as well. Inspection of the items showed solutions that match our theoretical constructs of teaching behaviour. The six components accounted for 55 % of the variance. The first component accounted for 41 % of the variance, the second for 4 %, the third for 3 %, and the remaining for about 2 % each. Regarding the reliability of items, the internal consistency estimates for all scales were above the satisfactory level, ranging from 0.81 (Clear instruction) to 0.98 (Activating learning) at the class level (see Table 1). Moreover, intra-class correlations between scales ranged between 0.18 (Clear instruction) and 0.50 (Activating learning), meaning that the student questionnaire scales could distinguish between teacher/class differences in teaching behaviour. Mean inter-scale correlations ranged from 0.71 (Classroom management) to 0.78 (Activating learning). This indicates that, although there was an overlap among aspects of teaching behaviour, the scales could measure distinct aspects of teaching behaviour as well. Again, a much larger portion of variance was explained by the first component than the remaining ones and high inter-scale correlations provided preliminary support for the unidimensional

Table 1 Reliability and descriptive statistics of observation and student questionnaire scales on teaching behaviour

Scale	Number of items		Cronbach α		Mean score		SD		Intra-class correlation		Mean inter-scale correlation	
	Obs	Que	Obs	Que	Obs	Que	Obs	Que	Obs	Que	Obs	Que
Safe and stimulating learning climate	4	12	0.80	0.96	3.20	2.89	0.60	0.57	0.49	0.43	0.59	0.74
Efficient classroom management	4	8	0.84	0.97	2.94	2.82	0.70	0.66	0.56	0.49	0.63	0.71
Clear instruction	7	11	0.88	0.81	2.92	2.80	0.66	0.45	0.53	0.18	0.69	0.75
Activating learning	7	17	0.81	0.98	2.68	2.70	0.60	0.65	0.37	0.50	0.69	0.78
Adaptation of teaching	4	6	0.72	0.93	2.05	2.80	0.80	0.62	0.39	0.47	0.60	0.72
Teaching–learning strategy	6	10	0.89	0.94	2.30	2.70	0.86	0.55	0.60	0.35	0.60	0.72

Obs Peer teacher data, *Que* student survey data

construct of teaching behaviour as measured by the student questionnaire. Additionally, a slightly-moderate correlation between results of observations and student perceptions (at the aggregated class level) was found ($r = 0.26$, $p < 0.05$).

Inspection of the mean scores between observation and student survey scales indicated that classrooms differed in the degree to which teaching behaviour was perceived by the trained observers and students (see Table 1). In general, observers seemed to have more favourable judgements with regard to Learning climate, Classroom management, and Clarity of instruction than students, while students appeared to have more positive judgements regarding Activating learning, Adaptation of teaching, and Teaching learning strategy than peer teachers. Together, observers seemed to have a more positive judgement regarding the less complex teaching behaviour than students, while students tended to have a more-positive judgement regarding the more-complex teaching behaviour compared to observers.

Rasch analysis

Because the six components of teaching behaviour based on observation and student surveys were highly correlated, we explored the construct representation of teaching behaviour in more detail using Rasch analysis with dichotomous variables. To test the unidimensionality of the teaching behaviour construct with dichotomous variables, we initially examined the scree plots based on results of tetrachoric correlations matrix using EFA (Demars 2010). The scree plots based on observation as well as student data appeared to show one dominant factor (see Fig. 1). This suggests that the unidimensionality of the teaching behaviour construct is reasonable. However, an examination of scree plots based on EFA results is a heuristic approach rather than a statistical testing procedure. To confirm the unidimensionality of the teaching behaviour construct, we performed Rasch modelling (see Table 2).

Results of the Rasch modelling for observation data revealed that the data were best represented by the one-subgroup solution (CAIC = 2541.61). This means that the ordinary unidimensional Rasch model fitted the data better than the mixed Rasch model (CAIC = 2644.35). The reliability estimate for the ordinary unidimensional Rasch model was 0.89, indicating a good reliability of the observation as a unidimensional measure (see Table 3). Additionally, all four fit statistics for the ordinary Rasch model indicated good fit

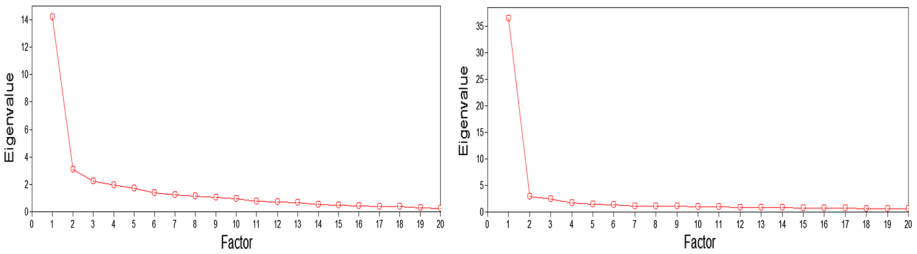

Fig. 1 Scree plots based on tetrachoric correlations for peer teacher ratings (*left*) and student perceptions (*right*) of teaching behaviour

with the data ($p > 0.05$). Hence, observation data held the property of a Rasch model sufficiently, suggesting that teaching behaviour can be regarded as a unidimensional latent construct.

For student data, the mixed Rasch model with a two-subgroups solution (CAIC = 43,461.51) appeared to fit the data better than the ordinary Rasch model (CAIC = 43,914.21). The relatively best solution, according to CAIC, was the two-subgroups solution because only a trivial change in the CAIC (ΔCAIC = 2.63) was visible from the two-subgroups to the three-subgroups solution. Besides, the BIC index increased from the two-subgroup to the three-subgroup solution. Hence, the two-subgroups solution was accepted as appropriate for the student data. The size in each subgroup was: Subgroup 1, 61 %; Subgroup 2, 39 %.

These findings suggest that student perception data were non-homogeneous as a whole, but could be classified into two Rasch-homogeneous subsamples. Nevertheless, Cressie-Read and Pearson χ^2 statistics were significant ($p = 0.00$), indicating that, although the mixed Rasch model fitted the data better, there still was room for improvement. Table 3 shows the mean teaching behaviour estimates, standard deviations, mean raw scores and reliabilities for each subgroup of the student ratings. The teaching behaviour estimates differed between subgroups, accounting for 2.53 for Subgroup 1 and −0.32 for Subgroup 2. Reliabilities appear to be good for both subgroups, ranging from 0.86 (Subgroup 1) to 0.94 (Subgroup 2). The inter-correlation between Subgroups 1 and 2 was 0.05, suggesting that the item difficulty orders were highly distinct between the subgroups.

Additional inspection of the results of Rasch modelling associated with individual preservice teachers' performance level of teaching behaviour (see Fig. 2) showed different patterns of performance levels, based on observations and student perceptions. According to trained observers, the performance level of the majority of preservice teachers was low (50 %), followed by high (31 %) and average (19 %), respectively. In contrast, student perceptions showed that the performance level of the majority of preservice teachers was high (49 %), followed by a rather similar proportion for low and average (26 % each).

Predictive quality of observations and student perceptions of teaching behaviour on student academic engagement.

Multilevel modelling revealed an important role of teaching behaviour for students' academic engagement as perceived by trained observers and students (see Table 4). Both observations and student perceptions of teaching behaviour could significantly predict students' academic engagement. However, the significant effect of student perceptions ($p < 0.01$) revealed a more powerful prediction than that of observations ($p < 0.05$). Student perceptions of teaching behaviour could explain about 13 % of the variance in students' academic engagement, while observations could explain only about 4 % of the

Table 2 Results of Rasch modelling for an ordinary unidimensional Rasch model and mixed Rasch model on teaching behaviour

Model	Observation data							Questionnaire data						
	LogL	N par	CAIC	LR	C-R	Pearson χ^2	FT	LogL	N par	CAIC	LR	C-R	Pearson χ^2	FT
Rasch model	−1183.29	33	2541.61	0.08	0.25	0.38	0.83	−21705.03	65	43,914.21	0.00	0.00	0.00	0.90
Mixed Rasch model														
Two subgroups	−1144.49	67	2644.35	0.03	0.40	0.43	0.10	−21,222.79	133	43,461.51	0.80	0.00	0.00	0.95
Three subgroups								−20,965.57	200	43,458.86	0.00	0.00	0.00	0.63

CAIC is fit statistic for between-model comparisons, LR, LR, C-R, and FT are fit statistics for within-model comparisons (Von Davier and Rost 1994)

LogL Loglikelihood, *Npar* number of parameters, *CAIC* consistent akaike information criterion, *LR* likelihood ratio, *C-R* Cressie-Read statistic, *FT* Freeman-Tukey statistic

Table 3 Descriptive statistics of the one-subgroup solution of observation data and the two-subgroup solutions of student survey data

Statistic	Observation data	Questionnaire data	
		Subgroup 1	Subgroup 2
Mean teaching behaviour	0.59	2.53	−0.32
SD teaching behaviour	1.64	1.45	1.39
Mean score	18.46	53.68	28.42
Reliability	0.89	0.86	0.94
Intercorrelation	NA	0.05	

NA Not applicable

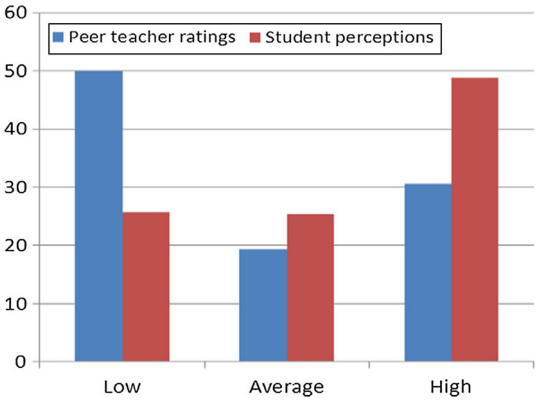

Fig. 2 Peer teacher ratings and student perceptions of teaching behaviour for three performance levels based on results of the best fitted Rasch model (percentages)

variance in engagement. Together, both observations and student perceptions of teaching behaviour could explain about 14 % of the variance in students' academic engagement.[4] The effects remained statistically significant even after adjusting for some controlled variables. This suggests that student perceptions of teaching behaviour were more predictive of their academic engagement than observations, although observations remained important as well. Model 4 also indicated that teaching subject had a negative effect on student engagement, which indicates that students in non-science classes (e.g. language, economy, sociology) reported lower levels of academic engagement compared to their peers in science classes (mathematics, biology, physics, chemistry).

Conclusions and discussion

Based on analyses of the construct representation of observations and student perceptions measuring the teaching behaviour of preservice teachers, we found that both instruments revealed a comparable factor structure in terms of the number of factors and the magnitudes of

[4] We also analysed a single-level model regarding the relationship between peer teacher ratings and student perceptions of teaching behaviour and students' academic engagement. As expected, we found larger effects in the single-level model compared with the multilevel model (student perceptions explained about 28 % of the variance, peer teacher ratings explained about 16 % of the variance, and together they explained about 33 % of the variance). When the data were structured hierarchically (i.e. students nested within classes), estimations based on the single-level model tended to overestimate the effects. Hence, the multi-level model was more appropriate.

Table 4 Results of multilevel analysis to explain variation in student academic engagement: parameter estimates

Variable	Academic engagement									
	Model 0		Model 1		Model 2		Model 3		Model 4	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed effects										
Intercept	3.00***	0.02	2.97***	0.03	2.87***	0.02	2.87***	0.03	2.80**	0.09
Observations			0.04*	0.02			0.02	0.02	0.02	0.02
Student perceptions					0.11***	0.01	0.11***	0.01	0.11***	0.01
Teaching subject									−0.16*	0.08
Teacher gender									0.08	0.06
Student gender									0.05	0.03
Random effects										
Class level	0.04	0.01	0.03	0.01	0.03	0.01	0.03	0.01	0.03	0.01
Student level	0.27	0.01	0.27	0.01	0.24	0.01	0.25	0.01	0.25	0.01
Deviance	3408.70	2075.69	2559.32	1571.98	1477.15					

* $p < 0.05$; *** $p < 0.01$

eigenvalues. The six-factor structure underlying one dimension of the teaching behaviour construct, found in both instruments, was consistent with the number of hypothesised scales of teaching behaviour. Furthermore, both instruments showed reliabilities at above satisfactory levels (student data: 0.81–0.98; Observation data: 0.72–0.89; see Table 1), were able to detect between-class differences with regard to teaching behaviour, and revealed that, although overlapping was inevitable, each factor measured a hypothesised scale of teaching behaviour. This preliminary result suggested that, in terms of classical test theory analysed heuristically, the construct structure of observations and student perceptions instruments was comparable. However, the low correlation between observations and student perceptions of teaching behaviour could be indicative of construct differences being measured.

Furthermore, Rasch analysis showed that the construct representation of observation data was different from that of student data. The teaching behaviour construct in observation data was best represented by an ordinary unidimensional Rasch model, suggesting that all items in the observation instrument could be ordered on a single latent continuum, and that this ordering of items was identical for all respondents. No structural differences were found in the observation data. Hence, the same unidimensional construct, namely, teaching behaviour, was measured uniformly and similar interpretations applied. Teaching behaviour appeared to be well measured in these data. This finding is consistent with the previous research that the ordinary Rasch model represents the observation instrument measuring teaching behaviour in the context of primary education (Van de Grift 2007) as well as secondary education (Van de Grift et al. 2013).

The Rasch model applied to student data as well, but the mixed Rasch model with two-subgroups solution fitted the data better than the more restrictive ordinary Rasch model and the mixed Rasch model with three-subgroups solution. In the mixed Rasch model with two-subgroups solution, the Rasch model held within each subgroup, but structural differences in item parameters between the two subgroups existed. This suggests that the construct representation of the student survey data was affected by latent subgroups, suggesting that the same construct was not measured uniformly for all respondents and the same interpretations did not apply. This implies that interpretations of teaching behaviour in traditional test use based on a single group is not appropriate.

A relatively different construct characteristic between observations and student perceptions indicated by the Rasch modelling results provides support for rather low agreement between the two measures of teaching behaviour. At the dimension level, the same theoretical construct of teaching behaviour was being measured from the perspective of trained observers and students. Observers observe teaching behaviour in a uniform way in relation to the ordering of item difficulties, while two subgroups of students perceived the rank order and item difficulties differently. A low agreement in the perceptions of teaching behaviour measured by observations and student perceptions suggests that, even when the latent construct is theoretically identical, it is less clear whether or not the same theoretical construct is being assessed from the perspective of observers and students. However, this does not necessarily mean that trained observers and secondary-school students are not able to recognise characteristics of classroom environments reflective of preservice teachers' teaching behaviour. Rather, both respondents generally seem to interpret the construct in question rather differently. Patton (1980) argued that differences in perspectives are advantageous and provide the opportunity for better understanding the complexity of classroom practices.

A rather moderate agreement in the construct representation found is probably attributable to the more positive perceptions of students than observers as was also indicated in a past study (Marsh 1984). Indeed, we found that students perceived that the majority of their teachers have a high level of teaching performance. By contrast, observers rated the majority of teachers as having a low level of teaching performance. Given that

both measures of teaching behaviour appear to measure the teaching behaviour of preservice teachers reliably, the slight moderate agreement seems to be a function of differences in observers and student perceptions, as was also shown in a past study (Lawrenz et al. 2003). Regardless of this, results from observations and student perceptions could be complementary in unravelling the complexity of teaching behaviour from different perspectives and suggest that the reality of the situation is not based on a single truth (Fraser 2012; Saljo 1997). Another possible explanation of a lack of convergence regarding the construct representation between the two measures might be attributable to differences at the item level. Although the two measures were developed in a rather identical way, the student questionnaire has more items than the observation instrument. Hence, it is reasonable to assume that the student questionnaire taps more comprehensive information about teaching behaviour compared with the observation instrument.

Regarding the relationships with student academic engagement, student perceptions of teaching behaviour was more predictive than observations. This finding is consistent with past studies associated with the superior predictive quality of student perceptions on student outcomes (De Jong and Westerhof 2001; Fraser 1991; Kunter and Baumert 2006; Maulana 2012a, b; Seidel and Shavelson 2007). This finding is also in line with the meta-analysis of Roorda et al. (2011), showing that associations between teaching behaviour and engagement are stronger when behaviour and engagement are measured using the same informants. Observations of teaching behaviour were significant predictors of student engagement as well. This suggests that the inclusion of observations is not only useful for assessing the performance of preservice teachers, but it also supports the additional significance of observations in examinations of classroom practices related to student outcomes. Consistent with this, Lawrenz et al. (2003) and Cornelius-White (2007) showed that both observer and student perceptions are effective measures of classroom practices in relation to student outcomes. Observations could also be an effective tool for feedback purposes leading to enhanced effective learning outcomes for preservice teachers (Lynch et al. 2012).

The finding that the mixed model fitted to the student data leads us to conclude that student perceptions of behaviour are not to be interpreted without caution. One could argue that teaching experience could explain this result. Inexperienced teachers are often more unclear and variable in their behaviour, which could compromise certain student perceptions. An alternative explanation could be that students' lack of knowledge and skills regarding teaching makes it impossible to recognise the quality of more-complex teaching behaviour, in the same way that novice teachers also fail to recognise this complexity (Berliner 1988). Some students might have more experience in teaching (teaching siblings, providing homework classes, or by providing sport instructions). These students might be responsible for the mixed Rasch model results. Further research is called for to enhance insight into the processes underlying student judgements. These underlying mechanisms have been shown to be related to students' academic engagement in this study. It seems to be relevant to unravel these psychological processes in order to help (student) teachers and researchers to interpret student evaluations and, more importantly, to comprehend how teachers affect their students.

Additionally, we found that observers seemed to have more-favourable judgements regarding learning climate, classroom management and clarity of instruction than students, while students appeared to have more positive judgements regarding activating learning, adaptation of teaching and teaching–learning strategies than peer teachers. More particularly, observers rated the adaptation of teaching and teaching–learning strategy significantly lower than students. More in-depth study is needed because reasons for these differences remain inconclusive. One possible explanation could be that observers tend to rate lower because of rather limited time to observe preservice teachers (based solely on one lesson). One might

argue that some behaviours have a higher chance of occurring on one occasion than another. Subsequently, this can be seen as a potential limitation of the current study. We acknowledge that assessing teaching by means of observations is a complex and costly approach. When possible, future researchers should try to incorporate more observation moments to get a more representative picture of teaching behaviour from the observer perspective. In the context of preservice teachers, they can use the observation instrument and student questionnaire as diagnostic tools to assess their own teaching and also that of their school mentors. An additional explanation can be sought in the ability of experienced teachers to recognise the complexity of teaching behaviour (Berliner 1988). The low ratings might reflect trained observers' ability to recognise readily the skills that are lacking. Information gathered from these tools is useful for their learning process in becoming effective teachers.

In conclusion, the present study contributes to the existing debate about potential differences in the construct representation of teaching behaviour measured by observations and student surveys by means of Rasch modelling. The Rasch approach has been useful for determining whether or not the same constructs are being measured from two different informants (student perceptions and observer ratings). Consistent with the idea of Cavanagh and Waugh (2011), Rasch measurement can be applied to learning environments research for assessing the quality of instruments more precisely than classical test theory. Additionally, the study supplements the body of knowledge regarding how best to examine teaching practices in terms of teaching behaviours that meet the educational initiatives for improving student outcomes. Good reliability and validity for both observations and student perceptions offer more alternatives for examining classroom practices, depending on the specific goals. For teacher education particularly, our study suggests that both instruments can be used effectively to support the teaching performance of preservice teachers. A relatively moderate agreement in the construct representation of teaching behaviour as perceived by observers and students offers ideas for researchers in the field of learning environments and teacher education for investigating different perspectives of teaching behaviour more closely. Differences in perspectives reflect a wealth of information about the complexity of classroom practices.

For an examination of links between teaching behaviour and student outcomes, student perceptions provide the most informative approach. Additionally, observations appear to be more appropriate for more in-depth examination of teaching behaviour by continuing the investigation of low-inference indicators. Hence, the decision about which measure to use for assessing classroom practices should depend on the specific goal in mind. Given that the student questionnaire used in the present study is relatively lengthy, future research would benefit from a more-economical version with comparable (or even better) psychometric qualities. Finally, it is our hope that this article will stimulate global discussion among researchers in the field of learning environments and teacher education regarding the examination of teaching skills across countries. Ideas for collaboration among countries regarding the international comparison of teaching skills are welcome.

Acknowledgments This study was part of a national research project about the relationship between the development of the quality of teaching behaviour of preservice teachers and the preparation route of teacher education in Dutch secondary education (http://www.rug.nl/lerarenopleiding/onderzoek/teachingTeacherEducation/inductie?lang0nl, http://www.rug.nl/lerarenopleiding/onderzoek/opleidenindeschool/index). This project was financed by the Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO, project number 411-09-802). NWO funds scientific research at Dutch universities and institutes.

Appendix 1

See Table 5.

Table 5 Examples of dimensions and the corresponding items of observation instrument

Dimension	High inference indicator		Low inference indicator	
	The teacher…	Rating	The teacher…	Observed (0 = no, 1 = yes)
Safe and stimulating learning environment	Shows respect for the pupils in behaviour and language use	1 2 3 4	Allows pupils to finish speaking	0 1
			Listens to what pupils have to say	0 1
			Makes no role-confirming remarks	0 1
	Ensures a relaxed atmosphere	1 2 3 4	Addresses the children in a positive manner	0 1
			Reacts with humour and stimulates humour	0 1
			Demonstrates warmth and empathy toward all pupils	0 1
Efficient classroom management	Ensures the lesson proceeds in an orderly manner	1 2 3 4	Intervenes timely and appropriately in case of disorder	0 1
			Safeguards the agreed rules and codes of conduct	0 1
	Uses the time for learning efficiently	1 2 3 4	Starts the lesson on time	0 1
			Does not keep pupils waiting	0 1
Clear instruction	Presents and explains the subject material in a clear manner	1 2 3 4	Gives staged instructions	0 1
			Poses questions which pupils can understand	0 1
	Gives clear explanation of how to use didactic aids and how to carry out assignments	1 2 3 4	Explains how lesson aims and assignments relate to each other	0 1
			Explains clearly which materials and sources can be used	0 1
Activating learning	Stimulates pupils to think about solutions	1 2 3 4	Shows pupils they can take towards a solution	0 1
			Shows learners how to consult sources and reference works	0 1
	Gives interactive instructions	1 2 3 4	Promotes the interaction between pupils	0 1
			Promotes the interaction between teacher and pupils	0 1

Table 5 continued

Dimension	High inference indicator		Low inference indicator	
	The teacher…	Rating	The teacher…	Observed (0 = no, 1 = yes)
Adaptation of teaching	Offers weaker learners extra study and instruction time	1 2 3 4	Gives weaker learner extra study time	0 1
			Gives weaker learners extra exercises/practices	0 1
	Adjusts instructions to relevant inter-learner differences	1 2 3 4	Gives additional instructions to small groups or individual pupils	0 1
			Does not simply focus on the average learner	0 1
Teaching learning strategy	Teaches pupils how to simplify complex problems	1 2 3 4	Teaches pupils how to break down complex problems into simpler ones	0 1
			Teaches pupils to order complex problems	0 1
	Teaches pupils to check solutions	1 2 3 4	Teaches pupils how to estimate outcomes	0 1
			Teaches pupils how to predict outcomes	0 1

Appendix 2

See Table 6.

Table 6 Examples of dimensions and the corresponding items of student questionnaire

Dimension	My teacher…	Rating
Safe and stimulating learning environment	Treats me with respect	1 2 3 4
	Ensures that I feel relaxed in class	1 2 3 4
Efficient classroom management	Applies clear rules	1 2 3 4
	Ensures that I use my time effectively	1 2 3 4
Clear instructions	Uses clear examples	1 2 3 4
	Explains everything clearly to me	1 2 3 4
Activating learning	Involves me in the lesson	1 2 3 4
	Ensures that I do my best	1 2 3 4
Adaptation of teaching	Connects to what I know or am capable of	1 2 3 4
	Checks whether I have understood the subject matter	1 2 3 4
Teaching learning strategy	Explains how I should study something	1 2 3 4
	Explains how I need to do things	1 2 3 4

References

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110–145). London: Sage.

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—Secondary. *School Psychology Review, 42*, 76–79.

Anderson, A., Hamilton, R., & Hattie, J. (2004). Classroom climate and motivated behaviour in secondary schools. *Learning Environments Research, 7*, 211–225.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of School Psychology, 44*, 427–445.

Berliner, D. C. (1988). *The development of expertise in pedagogy*. New Orleans: American Association of Colleges for Teacher Education.

Bozdogan, H. (1987). Model selection and Akaikes information criteria (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.

Cavanagh, R. F., & Romanoski, J. T. (2006). Rating scale instruments and measurement. *Learning Environment Research, 9*, 273–289.

Cavanagh, R. F., & Waugh, R. F. (2011). *Application of Rasch measurement in learning environments research*. Rotterdam: Sense Publishers.

Cornelius-White, J. (2007). Learner-centred teacher-student relationships are effective: A meta-analysis. *Review of Educational Research, 77*, 113–143.

Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.

De Jong, R., & Westerhof, K. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*, 51–85.

Demars, C. (2010). *Items response theory: Understanding statistics measurement*. New York: Oxford University Press.

Denzin, N. K. (1997). Triangulation in educational research. In J. P. Keeves (Ed.), *Educational research methodology and measurement: An international handbook* (pp. 318–322). New York: Pergamon.

Embretson, S. E. (2007). Mixed Rasch models for measurement in cognitive psychology. In M. T. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch model: Extensions and applications* (pp. 235–253). New York: Springer.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59*, 117–142.

Fraser, B. (1991). Two decades of classroom environment research. In H. J. Walberg (Ed.), *Educational environments: Evaluation, antecedents and consequences* (pp. 3–27). Elmsford: Pergamon.

Fraser, B. (2007). Classroom learning environments. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 103–124). Mahwah: Lawrence Erlbaum.

Fraser, B. (2012). Classroom learning environments: Retrospect, context, and prospect. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 1191–1239). New York: Springer.

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology, 95*, 148–162.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74*, 262–273.

Ko, J., & Sammons, P. (2013). *Effective teaching: A review of research and evidence*. CfBT Education Trust. http://cdn.cfbt.com/~/media/cfbtcorporate/files/research/2013/r-effective-teaching-2013.pdf.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*, 231–251.

Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*, 12–23.

Lasagabaster, D., & Sierrra, J. M. (2011). Classroom observation: Desirable conditions established by teachers. *European Journal of Teacher Education, 34*(4), 449–463.

Lawrenz, F., Huffman, D., & Robey, J. (2003). Relationships among student, teacher and observer perceptions of science classrooms and student achievement. *International Journal of Science Education, 25*(3), 409–420.

Lynch, R., McNamara, P. M., & Seery, N. (2012). Promoting deep learning in a teacher education program through self- and peer-assessment and feedback. *European Journal of Teacher Education, 35*(2), 179–197.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707–754.

Maulana, R. (2012). *Teacher-student relationships during the first year of secondary education*. Groningen: University of Groningen.

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teaching behavior: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*, 169–194.

Maulana, R., Opdenakker, M.-C., den Brok, P., & Bosker, R. (2011). Teacher-student interpersonal relationships in Indonesian secondary education: Profiles and importance to student motivation. *Asia Pacific Journal of Education, 31*(1), 33–49.

Maulana, R., Opdenakker, M.-C., den Brok, P., & Bosker, R. (2012a). Teacher–student interpersonal relationships in Indonesian lower secondary education: Teacher and student perceptions. *Learning Environments Research, 15*, 251–271.

Maulana, R., Opdenakker, M.-C., Stroet, K., & Bosker, R. (2012b). Observed lesson structure during the first year of secondary education: Exploration of change and link with academic engagement. *Teaching and Teacher Education, 28*(6), 835–850.

Ministerie van Onderwijs, Cultuur en Wetenschap. (2012). *Nota werken in het onderwijs 2012*. http://www.rijksoverheid.nl/documenten-en-publicaties/brochures/2011/09/23/nota-werken-in-het-onderwijs.html. Accessed 5 March 2013.

Muthen, L. K., & Muthen, B. O. (1999). *Mplus users' guide: The comprehensive modeling program for applied researchers*. Los Angeles: Muthen & Muthen.

OECD. (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science*. Paris: OECD.

Opdenakker, M.-C., Maulana, R., & den Brok, P. (2012). Teacher–student interpersonal relationships and academic motivation within one school year: Developmental changes and linkage. *School Effectiveness and School Improvement, 23*, 95–119.

Patrick, H., & Ryan, M. (2007). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology, 99*, 83–98.

Patton, M. Q. (1980). *Qualitative evaluation methods*. Beverly Hills: Sage.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*, 109–119.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). *MLwiN Version 2.0*. Bristol: Centre for Multilevel Modelling, University of Bristol.

Riggin, L. J. (1997). Advances in mixed-method evaluation: A synthesis and comment. *New Directions for Evaluation, 74*, 87–94.

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher-student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*, 493–529.

Sabers, D., Cushing, K., & Berliner, D. C. (1991). Differences among expert, novice, and postulant teachers in a task characterized by simultaneity, multidimensionality and immediacy. *American Educational Research Journal, 28*, 63–88.

Saljo, R. (1997). Self-report in educational research. In J. P. Keeves (Ed.), *Educational research methodology and measurement: An international handbook* (pp. 101–105). New York: Pergamon.

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London: Office for Standards in Education.

Sammons, P. M., & Ko, J. (2008). *Using systematic classroom observation schedules to investigate effective teaching: Overview of quantitative findings, Effective Classroom Practice (ECP),* ESRC Project Report. Nottingham: School of Education, University of Nottingham.

Scantlebury, K., Boone, W. J., Kahle, J. B., & Fraser, B. J. (2001). Design, validation, and use of an evaluation instrument for monitoring systematic reform. *Journal of Research in Science Teaching, 38*, 646–662.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Shuell, T. J. (1996). Teaching and learning in a classroom context. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 726–764). New York: Macmillan.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modelling, 3*, 25–40.

Spooren, P., Mortelmans, D., & Thijssen, P. (2012). 'Content' versus 'style': Acquiescence in student evaluation of teaching. *British Journal of Educational Research, 38*(1), 3–21.

Tu, S., & Xu, L. (2012). A theoretical investigation of several model selection criteria for dimensionality reduction. *Pattern Recognition Letters, 33*, 1117–1126.

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research, 49*(2), 127–152.

Van de Grift, W., & Van der Wal. (2010). The measurement of teaching skills with the Rasch model. Paper presented at the International Congress on School Effectiveness and Improvement, Malaysia.

Van de Grift, W., & Helms-Lorenz, M. (2012). *Vaardigheid en ervaring van leraren [Skills and experience of teachers]*. Groningen: Department of Teacher Education, University of Groningen.

Van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2013). *Observeren van leraren in opleiding [Observed teaching skills of pre-service teachers]*. Groningen: Department of Teacher Education, University of Groningen.

Van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*, 150–159.

Van de Grift, W., Houtveen, T., & Vermeulen, C. (1997). Instructional climate in Dutch secondary education. *School Effectiveness and School Improvement, 8*(4), 449–462.

Van de Grift, W. J. C. M., Van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs. *Pedagogische Studiën, 88*, 416–432.

Virtanen, T. E., Lerkkanen, M.-J., Poikkeus, A.-M., & Kuorelahti, M. (2013). The relationship between classroom quality and students' engagement in secondary school. *Educational Psychology,*. doi:10.1080/01443410.2013.822961.

Von Davier, M. (1994). *WINMIRA: A program system for analyses with the Rasch model, with the latent class analysis and with the mixed Rasch model*. Kiel: Institute for Science Education (IPN).

Von Davier, M., & Rost, J. (1994). Self-monitoring—A class variable? In J. Rost, & R. Langeheine (Eds.), *Applications of latent trait and latent class models*. Proceedings of the IPN Symposium in Sankelmark 1994.

Woolley, M. E., & Bowen, G. (2007). In the context of risk: Supportive adults and the school engagement of middle school students. *Family Relations, 56*, 92–104.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: An alternative approaches and practical guidelines*. White Plains: Longman.

Wubbels, T., & Brekelmans, M. (2005). Two decades of research on teacher-student relationships in class. *International Journal of Educational Research, 43*(1–2), 6–24.

Wubbels, T., & Brekelmans, M. (2012). Teacher-student relationships in the classroom. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 1241–1256). New York: Springer.